FEATURE AND SCORE NORMALIZATION FOR SPEAKER VERIFICATION OF CELLULAR DATA *

Claude Barras and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France

{barras,gauvain}@limsi.fr

ABSTRACT

This paper presents some experiments with feature and score normalization for text-independent speaker verification of cellular data. The speaker verification system is based on cepstral features and Gaussian mixture models with 1024 components. The following methods, which have been proposed for feature and score normalization, are reviewed and evaluated on cellular data: cepstral mean subtraction (CMS), variance normalization, feature warping, T-norm, Z-norm and the cohort method. We found that the combination of feature warping and T-norm gives the best results on the NIST 2002 test data (for the one-speaker detection task). Compared to a baseline system using both CMS and variance normalization and achieving a 0.410 minimal decision cost function (DCF), feature warping and T-norm respectively bring 8% and 12% relative reductions, whereas the combination of both techniques yields a 22% relative reduction, reaching a DCF of 0.320. This result approaches the state-of-the-art performance level obtained for speaker verification with land-line telephone speech.

1. INTRODUCTION

Over the last ten years, Gaussian mixture models (GMM) for the modeling of speaker spectral characteristics has become the dominant approach for speaker verification systems which use untranscribed training data [1].

Efficient normalization of the acoustic features and of the detection score have been recently proposed for improving a standard GMM-based system. For example, it has been reported in [4] that feature warping, which consists of mapping the observed short-time distribution of the acoustic features to a normal distribution, outperforms standard cepstral mean subtraction. Another example is the T-norm method [5] which uses the statistics of the scores of a cohort of impostor speakers to normalize the target speaker score.

In this paper the following normalization methods are reviewed and evaluated on cellular data: cepstral mean subtraction (CMS), variance normalization, feature warping, T-norm, Z-norm and the cohort method. Results are reported on the NIST 2002 test data for the one-speaker detection task [6].

In the next section we describe the experimental conditions and the baseline system. In Section 3 we review the various normalization methods. The experimental results are discussed in Section 4.

*This work was partially financed by the French Ministry of Defense

2. EXPERIMENTAL SETUP

In this section, we describe the NIST one-speaker detection task, the corpora used to carried out the experiments, and the baseline speaker verification system.

Corpus and task

The speaker recognition experiments were conducted on cellular telephone conversational speech from the Switchboard corpus. This data was used by NIST for the 2002 one-speaker detection task [7]. Given a speech segment of about 30 seconds, the goal is to to decide whether this segment was spoken by a specific target speaker or not. For each of 330 target speakers (139 males and 191 females), two minutes of untranscribed, concatenated speech is available for training the target model. Overall 3570 test segments (1442 males and 2128 females), mainly lasting between 15 and 45 seconds, have to be scored against roughly 10 gender-matching impostors and against the true speaker. The gender of the target speaker is known.

We made use of the cellular data from the NIST 2001 evaluation in order to train background models and to gather the needed statistics for some of the detection score normalization methods. This data includes files from 60 development speakers (2 minutes of speech for each of 38 males and 22 females) which are used to train the background models, and files from 174 target speakers (2 minutes of speech for each of 74 males and 100 females) used as impostor data, plus some of the 2038 evaluation test segments also used as impostor data.

Baseline system

PLP-like features are extracted from the speech signal every 10ms using a 30ms window. The feature vector estimated on the 0-3.8kHz bandwidth is comprised of 15 MEL-PLP cepstrum coefficients, 15 delta coefficients plus the delta energy, for a total of 31 features. Cepstral mean subtraction and variance normalization are applied to each speech file during training and testing. This front-end departs slightly from the one used in LIMSI speech recognition systems [11]. For speaker recognition we use 15 cepstral coefficients rather than 12, since higher order cepstral coefficients are known to carry some speaker information, and 15 was found to be the optimal on our development data. However the gain over 12 coefficients is quite small. Another difference concerns the second-order difference cepstral coefficients which were not found to be effective for speaker recognition [6].

For each target speaker, a speaker-specific GMM with diagonal covariance matrices was trained via maximum a posteriori (MAP) adaptation [8] of the Gaussian means of the matching gender background model using 5 iterations of the EM algorithm. Each of

the two gender-dependent background models includes 1024 Gaussians. These two models were trained on a total of about 2 hours of data from the 60 development speakers.

For each verification test, i.e.m a pair of a test segment and a target speaker, the test segment is scored against both the target model and the background model matching the target gender, ignoring low energy frames (about 10%). For a given test segment X and a target model λ, the decision score $S(X,\lambda)$ is a log-likelihood ratio,

$$S(X,\lambda) = \log f'(X|\lambda) - \log f'(X|R)$$

where $f'(X|\lambda)$ is the normalized likelihood of the speech segment (of length $L(X)$) for a given model, i.e., $f'(X|\lambda) = f(X|\lambda)^{1/L(X)}$, and $f'(X|R)$ is the normalized likelihood of the gender-matching background model.

Performance measure

A speaker detection system is subject to two kinds of errors, i.e. missed detections and false alarms. The primary performance measure for the NIST speaker detection task is the detection cost function (DCF) defined as a weighted sum of both error probabilities, the normalized cost taking the following form (see [7]) $C_{Norm} = P_{Miss} + 9.9 \times P_{FalseAlarm}$. For all results, we report the minimal DCF value obtained a posteriori for the best possible detection threshold. However this operating point favors false alarms, so the equal error-rate (EER) is used as an alternative performance measure. We are also using the Detection Error Tradeoff (DET) curves. The DET curve is comparable to the Receiver Operating Characteristics curve but the use of non-linear axis which results in a linear curve for a normal distribution improves its readability [9].

3. NORMALIZATION METHODS

This section reviews the various normalization methods tested in this work, i.e., cepstral mean normalization, variance normalization, and feature warping for feature normalization, and the cohort model, Z-norm and T-norm for score normalization.

Feature normalization

Speaker recognition systems generally make use of acoustic front-ends very similar to those used in speech recognition: signal band-limiting, cepstral feature extraction, feature normalization, cepstral and energy derivatives. Cepstral feature normalization often consists of cepstral mean subtraction (CMS) performed over the entire file. This reduces stationary convolution noises due to the channel. CMS is sometimes supplemented, as in our baseline, with variance normalization. Recently, other approaches have been successfully applied to feature normalization for speaker recognition, mainly feature warping [4] and short-time Gaussianization [10]. Feature warping consists of mapping the observed cepstral feature distribution to a normal distribution over a sliding window, the various cepstral coefficients being processed in parallel streams. Let r_t be the rank of a feature within a N sample window centered around time t, its warped value w_t is estimated by solving numerically the following equation:

$$\frac{r_t - 1/2}{N} = \int_{-\infty}^{w_t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

Feature warping has been shown to outperform standard normalization techniques [4]. Short-time Gaussianization is similar but applies a linear transformation to the features before mapping them to a normal distribution; this linear transformation which can be estimated by the EM algorithm, makes the resulting features better suited to diagonal covariance GMMs. The linear transformation can also be optimized separately for male and female speakers. Short-time Gaussianization was shown to perform better than feature warping for low false-alarm rates [10].

Score normalization

The baseline system score consists of the log-likelihood ratio between the target speaker and the reference background model matching the target gender. This approach is efficient but it is also quite fast as only two likelihoods values need to be computed [1]. A classical alternative for the estimation of the reference likelihood is to use a set of impostor models, or cohort. Following previous work of LIMSI on speaker recognition [3] we combined the impostor scores in the following way:

$$S_{cohort}(X,\lambda) = \log f'(X|\lambda)^{\gamma} - \log \sum_{R \in Cohort} f'(X|R)^{\gamma}$$

where γ is a scaling parameter needed to compensate for independency assumptions, which needs to be optimized on some development data. In our experiments, the gender of the test segment was automatically determined based on the best scoring male or female background model, and only cohort speakers from the same gender are considered. Further reduction of the cohort was optionally performed by keeping only the top best scoring models.

The other category of score normalization methods consist of normalizing the distribution of the scores. Here we consider the Z-norm and T-norm methods which have proven to be quite efficient [5]. The Z-norm method normalizes the score distribution using target-specific statistics:

$$S_{znorm}(X,\lambda) = \frac{S(X,\lambda) - \mu_\lambda}{\sigma_\lambda}$$

where μ_λ and σ_λ are respectively the mean and standard deviation of the scores $S(Y_i,\lambda)$ of the target speaker against a set of impostor test segments Y_i. This normalization was originally proposed for joint handset and speaker normalization [1], but it is used here for speaker normalization only since handset compensation is not relevant for cellular speech. The T-norm method extends the standard cohort approach to score distribution scaling [5]. Each test segment is scored against a set of impostor cohort models R_j, and the likelihood of the test segment given the target speaker is normalized according to the mean μ_X and standard deviation σ_X of the likelihoods $f'(X|R_j)$:

$$S_{tnorm}(X,\lambda) = \frac{\log f'(X|\lambda) - \mu_X}{\sigma_X}$$

As was done for the cohort scoring, only the impostor targets of the same gender as the test segment, or a fraction of them, are used for computing μ_X and σ_X. At the opposite of Z-norm, it is possible to conduct the normalization directly using the likelihoods $f'(X|\lambda)$ rather than the scores $S(X,\lambda)$, since the log-likelihood of the background model is a constant shared by all tests and is thus canceled by mean subtraction.

4. EXPERIMENTAL RESULTS

Experimental results obtained with the various normalization methods are given bellow.

Nb. Gaussians	τ	*min. DCF*	*EER (%)*
16	100	0.638	17.2
32	75	0.578	15.1
64	50	0.529	13.3
128	25	0.490	11.9
256	20	0.447	10.8
512	15	0.424	10.0
baseline: 1024	10	0.410	9.9

Table 1: System performance for GMMs with increasing numbers of Gaussians, using a matching MAP adaptation weight τ.

MAP adaptation	*min. DCF*	*EER (%)*
mean only	0.410	9.9
mean+weight	0.434	10.7
mean+weight+variance	0.540	12.5

Table 2: Impact of MAP adaptation of Gaussian mean, weight and variance over speaker detection performance.

Baseline performance

Gaussian means of the target models are adapted from the reference models using MAP adaptation [8]. Evidently the prior weight (τ) needs to be tuned depending of the model complexity, ranging to about $\tau = 100$ for small model with 16 Gaussians to about $\tau = 10$ for system with 1024 Gaussians per gender model. System performances with 16 to 1024 Gaussians are reported in Table 1. The baseline system with 1024 Gaussians has minimal DCF of 0.410 and an ERR of 9.9%. As can be seen in Table 2, adapting the Gaussian weights or variances is not helpful, this was already in [5]. Concerning the number of EM iterations for the adaptation, at least 4 iterations are needed when adapting the Gaussian weights or variances in addition to the Gaussian means, but 2 iterations are enough when only adapting the Gaussian means.

Feature normalization

Here we analyse the results with the following feature normalization methods: CMS, CMS plus variance normalization, short-term variance normalization and feature warping. For the first two methods, the normalization is applied over the entire file. Performances reported in Table 3 show that these two methods have very similar EER, but normalization of the variance improves the DCF. Feature warping was found to be especially efficient when applied over a sliding window of about 3 seconds, which is consistent with the results reported in [4, 10]. We also tested mean, mean plus variance and variance-only normalization over a sliding window; but we only report the best solution corresponding to a short-term variance normalization on a 3 second window. Table 3 gives the minimal DCF and EER for the most interesting conditions and Figure 1 contains the corresponding DET curves.

A significant part of the gain observed with feature warping seems to be due to the short-term variance normalization[1]. It may be the case that cellular speech, as used in our experiments, is especially subject to varying additive noises, which are better addressed by a short-term variance normalization.

Globally, feature warping brings about a 10% relative reduction of the minimal DCF, and a 15% relative reduction of the EER compared to a standard CMS solution.

Feature normalization	*min. DCF*	*EER (%)*
none	0.544	13.3
mean	0.422	10.0
mean and variance	0.410	9.9
short-term variance	0.394	8.9
feature warping	0.378	8.5

Table 3: Speaker detection performances for mean, mean and variance, short-term variance and feature warping normalizations.

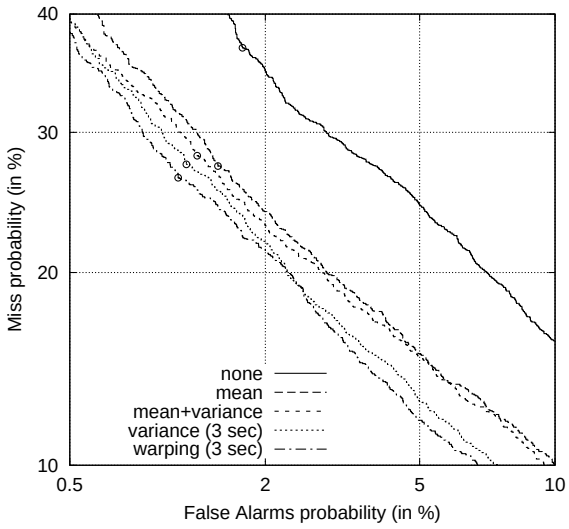

Figure 1: DET curves for the feature normalizations of Table 3. Circles are drawn at minimal DCF operating point.

Score normalization	*min. DCF*	*EER (%)*
baseline	0.410	9.9
cohort (all set, γ=1)	0.437	12.4
cohort (all set, γ=6)	0.376	11.0
cohort (top 6, γ=0.5)	0.363	10.9
T-norm (all set)	0.365	10.4
T-norm (top 80%)	0.360	10.4
Z-norm	0.405	9.6

Table 4: System performance for cohort normalization, Z-norm and T-norm over baseline system.

Score normalization

Cohort normalization was done using impostor models from 60 male and 86 female speakers[2] trained in a similar manner as the target models. A simple sum of all cohort scores (i.e., $\gamma = 1$) significantly decreases the performance over the baseline system as shown in Table 4. Using $\gamma = 6$ results in a much better DCF (a high value for γ emphasizes the highest impostor likelihoods). The best result is obtained by keeping only the 6 highest impostor likelihoods with $\gamma = 0.5$, thus including in the reference likelihood only those speakers who are the most similar to the target. It should be noted however that in terms of ERR the cohort method gives less good results than the baseline setup.

[1] Feature warping was also applied to the energy, which was not the case for the mean and variance normalizations and reduced the minimal DCF from 0.384 to 0.378 but had no impact on the EER.

[2] The speakers still present in 2002 evaluation were discarded from the initial 174 speakers set.

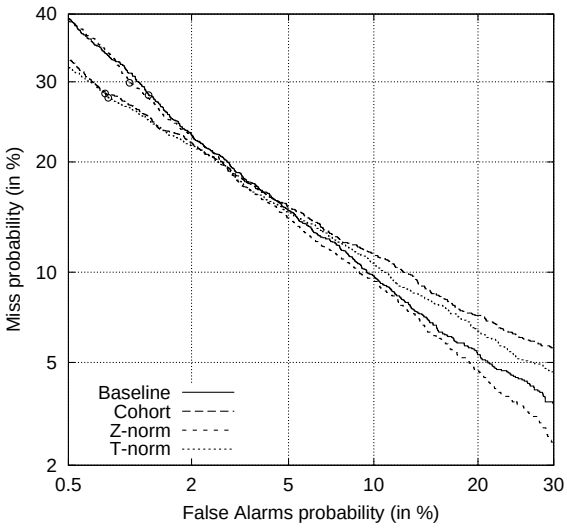

Figure 2: DET curves for cohort score normalization (using rhe 6 nearest impostor speakers), Z-norm (using 150 impostor tests) and T-norm (using 80% of impostor set).

T-norm gives better results than the optimized cohort method, the optimal DCF being reached by discarding the 20% most distant impostor speakers. The DET curves shown in Figure 2 reveal that T-norm performs better than the cohort method at all operating points. Both are better than the baseline in the area of minimum DCF, but this is reversed at low missed rates. The Z-norm parameters were estimated using 150 impostor test segments per gender. It can be seen in Table 4 that Z-norm gives a slight improvement over the baseline. The DET curves show that that Z-norm outperforms the baseline and the T-norm for low miss rates. Globally, we observe that T-norm and Z-norm affect DET curve by rotating it in opposite directions.

Combination of feature and score normalization

Table 5 summarizes the results obtained by combining feature warping and T-norm as extensions to the baseline system. Starting with a minimal DCF of 0.410 for the baseline system, feature warping and T-norm bring respectively 8% and 12% relative reductions of the cost, and the combination of both reduces the DCF by 22% reaching a minimal DCF of 0.320.

5. CONCLUSION

In this paper we have studied some feature and score normalization methods for speaker verification on cellular data. Experimental results were obtained for CMS and variance normalization, feature warping, T-norm, Z-norm and the cohort method. Applied to cellular telephone conversational speech, we observed that the main impact of feature warping is due to local variance normalization, and that it improves the DET curve over a large range of operating points. The T-norm score normalization extends the classical cohort approach by scaling the score distribution with the standard deviation of the impostor scores, which significantly improves system performance at low false alarm rates; however when favoring low miss rates other normalization methods like Z-norm are more adequate. Feature warping and T-norm seem to be independent since their effect is cumulative, resulting in more than a 20% relative reduction of the minimal DCF from 0.410 to 0.320. Such a system is approaching performance observed for land-line telephone speech under comparable conditions (i.e., 2 minutes of speech for training and 15-45 seconds of speech for test).

System	*min. DCF*	*EER (%)*
baseline	0.410	9.9
feature warping	0.378	8.5
T-norm	0.360	10.4
feature warping + T-norm	0.320	8.5

Table 5: System performance obtained for the combination of feature normalization (feature warping) and score normalization (T-norm).

REFERENCES

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] L. Lamel and J.L. Gauvain, "Speaker recognition with the Switchboard corpus," in *Proc. ICASSP*, Munich, Apr. 1997, vol. II, pp. 1067–1070.

[3] L. Lamel and J.L. Gauvain, "Speaker verification over the telephone," *Speech Communication*, vol. 31, pp. 141–154, 2000.

[4] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, June 2001.

[5] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[6] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[7] "The NIST year 2002 speaker recognition evaluation plan," 2002, http://www.nist.gov/speech/tests/spk/2002/doc/.

[8] J.-L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[9] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997.

[10] B. Xiang, U. Chaudhari, J. Navrátil, G. Ramaswamy, and R. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. ICASSP*, 2002, vol. 1, pp. 681–684.

[11] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.